ABSTRACT

Summary: We initiated the Predictive Toxicology Challenge (PTC) to stimulate the development of advanced SAR techniques for predictive toxicology models. The goal of this challenge is to predict the rodent carcinogenicity of new compounds based on the experimental results of the US National Toxicology Program (NTP). Submissions will be evaluated on quantitative and qualitative scales to select the most predictive models and those with the highest toxicological relevance.

Availability: http://www.informatik.uni-freiburg.de/~ml/ptc/

Contact: helma@informatik.uni-freiburg.de

BACKGROUND

Prevention of environmentally-induced cancers is a health issue of unquestionable importance. Almost every sphere of human activity in an industrialized society faces potential chemical hazards of some form. It is estimated that nearly 100,000 chemicals are in use in large amounts every day. A further 500–1000 are added every year. Only a small fraction of these chemicals have been evaluated for toxic effects like carcinogenicity. The US National Toxicology Program (NTP) http://ntp-server.niehs.nih.gov/contributes to this enterprise by conducting standardized chemical bioassays—exposure of rodents (mice and rats) to a range of chemicals—to help identify substances that may have carcinogenic effects on humans. However, obtaining empirical evidence from such bioassays is expensive and usually too slow to cope with the number of chemicals that can result in adverse effects on human exposure. This has resulted in an urgent need for carcinogenicity models based on chemical structures and properties alone. It is envisaged that such models would:

- generate reliable toxicity predictions for chemicals;
- enable low cost identification of hazardous chemicals; and
- refine and reduce the reliance on the use of large numbers of laboratory animals (Bristol et al., 1996)

The outcome of the bioassays conducted by the NTP has resulted in a large (by toxicological standards) database of compounds classified as carcinogens or otherwise. Predicting the outcome of these tests using chemical structure (and related information) presents a formidable test for techniques concerned with knowledge discovery from databases.

The present PTC is the successor of an earlier one (Srinivasan et al., 1999) conducted within the Machine Learning community and tries to address some of its shortcomings. Both competitions were inspired by the NIEHS Predictive-Toxicology Evaluation Project (Bristol et al., 1996), which involved contributions from human experts, expert systems, experimental techniques and data driven methods. It is similar in form to the CASP protein structure prediction meetings (http://PredictionCenter.llnl.gov/).

THE PREDICTIVE TOXICOLOGY CHALLENGE (PTC) 2000–2001

The Challenge is to obtain models that predict the outcome of biological tests for the carcinogenicity of chemicals using information related to chemical structure only. As rats and mice differ substantially in their response to chemical carcinogens we will require separate models for each species.

The stages of the Challenge are described below:

Data engineering (September 1, 2000–March 1, 2001)

This phase of the Challenge will be concerned with:

1. calculation of chemical descriptors and feature construction. The initial chemical structures will be augmented by new chemical and structural features/background relations; and
2. data cleaning. Any errors in the data identified by participants will be corrected.

We encourage chemists, toxicologists and developers of programs for feature construction to participate in this phase by sending us new and potentially useful descriptors.
for the chemicals involved. To minimize bias during model construction, we will specify the rules for constructing these descriptors and the requirements for accompanying documentation. Submissions made here will be examined for toxicological relevance during the evaluation stage (cf. Section Model evaluation).

Model construction (March 2, 2001–June 1, 2001)

At the start of this phase, all input data will be fixed to ensure a level playing field for all model developers. For each of the modeling tasks mentioned in the Challenge section, we will require submissions to provide points specifying a Receiver Operating Characteristic (ROC) curve. This will allow a cost-sensitive assessment of models in the evaluation stage (cf. Section Model evaluation). In due course, we will specify the complete requirements of submissions.

Model evaluation (June 2, 2001–August 15, 2001)

We are in the process of obtaining an independent database for use during the evaluation stage. From the ROC curves of all submissions we expect to identify a subset of models that are ‘optimal’ for different cost conditions. Of these we will:

1. identify models that are particularly relevant to toxicology; and
2. identify any contributions made during the data engineering (cf. Section Data engineering) stage that were particularly informative

This will be a result of a co-operative exercise between the developers of the ‘optimal’ models and our toxicologists (currently C. Helma, University of Freiburg, Germany; and if possible, D.W. Bristol of the NIEHS, USA).

Dissemination (September, 2001)

A special workshop will be held on the Challenge as part of the ECML/PKDD Conferences at the University of Freiburg in September, 2001. This workshop will provide all participants with the opportunity to present their work. At this point, we expect to be in a position to decide if the submissions warrant a special journal publication.

INITIAL DATA

Initial data are already available at the PTC homepage (cf. below). At present, it contains the molecular structures of the NTP dataset in various formats and a summary of the rodent carcinogenicity classifications extracted from the NTP Full Report. A summary of bugs and inconsistencies found in these datasets and cleaned versions of the datasets will be added soon.

SUBMISSIONS

Contributions may be submitted by anonymous ftp to ftp://helma.informatik.uni-freiburg.de/incoming/. An accompanying email to helma@informatik.uni-freiburg.de should contain a description of the submitted contents.

HOMEPAGE

Detailed instructions for participation and the datasets can be obtained from the PTC homepage http://www.informatik.uni-freiburg.de/~ml/ptc/. This page will be constantly updated during the course of the challenge. Questions concerning the PTC can be addressed to helma@informatik.uni-freiburg.de.

REFERENCES
